分類語彙表番号を付与した『現代日本語書き言葉均衡コーパス』の書籍・新聞・雑誌データ

《資料・情報》

分類語彙表番号を付与した『現代日本語書き言葉均衡コーパス』の書籍・新聞・雑誌データ

加藤祥・浅原正幸・山崎誠

1. はじめに

著者らは，『現代日本語書き言葉均衡コーパス』（Maekawa et al, 2014）（以下 BCCWJ）の書籍・新聞・雑誌データに『分類語彙表増補改訂版』（2004）の分類語彙表番号を付与し，文脈的な意味分類によってコーパスを調査することを可能にした。

自然言語処理の分野では，語義曖昧性解消のタスクの学習・評価データとして様々な語義タグつきデータが整備されてきた。日本語では，古くは新聞記事を対象とした EDR コーパスや RWCP コーパスなどが国語辞典に基づく語義タグを付与していた。また，BCCWJ の一部に対しても，岩波国語辞典の語義が付与され，SemEval-2010 Japanese WSD Task（Okumura et. al. 2011）では，日本語の語義曖昧性解消の基礎データとして用いられてきた。シソーラスに基づくデータとしては，日本語ワードネットに基づく語義タグ付きコーパス（Bond et. al. 2012）が整備されたが，英語データを翻訳したものであった。代表性をもつ自然な日本語コーパスに対するシソーラスに基づく語義つきデータは，これまで管見の限り作成されていなかった。

意味的な情報を付与したコーパスは，語義の調査にはもちろん，意味による文章特徴の分析，比喩のように表現と語の意味に差がある場合などの分析にも有用である。本稿は，情報付与基準と基礎情報とともに，本データの公開[注1]について報告する。

2. BCCWJ への分類語彙表番号付与

2.1 作業対象

BCCWJ コアデータに含まれる書籍サンプル（PB），新聞サンプル（PN），雑誌サンプル（PM）のそれぞれの一部である 347,094 語を対象とし，自立語[注2]の 182,166 語に分類語彙表番号を付与した。本作業の対象範囲は，他情報との重ね合わせの便から，BCCWJ コアデータサンプルのアノテーション優先順序[注3]に従い，優先順序が高いサンプルを選択している。作業にあたっては，UniDic 語彙素番号（小木曽・中村，2014）と分類語彙表番号を人手で対応させたデータ[注4]（Kondo et al., 2018）を用い，自立語について BCCWJ の短単位ごとに，対応可能性のある分類語彙表番号の 5 桁目まで（**表 1** 参照）を列挙した。

表 1 分類番号の構造（例：この（分類番号：3.1010））

類	部 門	中項目	分類項目
相（3）	関係（.1）	真偽（.10）	こそあど（.1010）

2.2 作業基準

作業者は，分類語彙表番号の選択肢から，文脈的に該当する意味分類を選択した。作業は，各部分を分割して複数名の作業者のうち担当した 1 名が行ったが，質疑と作業方針を常時共有した。このほか，個別的に判定が困難であった場合は作業者が備考を付与して報告し，著者が統括的な整理を行った。なお，作業担当者によって判定揺れの生じた一部の多義語については，数十名規模の多数決判定[注5]を試みている。

分類語彙表掲載語は，単義の場合もあるが，最大 11 義の多義もある。UniDic- 分類語彙表データにマッチした自立語の複数選択肢の列挙された短単位は約半数の 44.8％であった。文脈上意味的に該当する分類が分類語彙表にない場合として，語彙素番号がなく UniDic に未登録の語であるときと，UniDic の語彙素番号に対応する分類語彙表番号がない未登録の語義であるときがある。そこで，選択できる分類語彙表番号がなく新規番号が必要となる場合には，作業者の判断で新たに適切な番号を付与した。但し，最小限の文脈に依拠した意味とし，比喩的・慣用的な表現などは語源的な意味とする。内容に即した意味は，長単位[注6]で対応する。

語彙素と分類語彙表の UniDic の語彙素に対応する分類語彙表番号がない例としては，たとえば「咎め立てる」のように語彙素はあっても分類語彙表では「咎める」「咎め立てする」の見出し語しかないといった不整合の場合のほか，未知語，固有名詞，略語などが多くある。

未知語には「カム」「トゥゲザー」「リトル」のような外来語，固有名詞には「キャンティ」「四日市」のような地名，略語には「委（委員会）」「大（大学）」などのほか，2 語以上が組み合わさった「メルマガ」「厚労」「AP」「IH」などの例が散見される。このような場合，それぞれの短単位の意味に相当する意味分類を人手で記入する。たとえば，「厚労」は「厚生」「労働」，「自民」は「自由」「民主」のそれぞれの短単位に相当する複数の分類語彙表番号を付与した。掛詞やダジャレなど，一短単位について複数の意味が読み取れる場合にも，同様に複数の意味について分類語彙表番号を付与した。

なお，「一個」「一口」のように UniDic に短単位での登録がある語は，その単位での分類語彙表番号候補が挙がる。文脈上，「一口二口食べる」のように「一（数詞）」「口（助数詞）」として別の短単位と読むことが適切な場合であっても，「一口」を一短単位として「一口に言う」のような言語活動（.3100）や「一口乗る」のような奉仕（.3541）の分類語彙表番号が挙がるためである。このような場合は，「一口」が一短単位とされていても，「一」「口」を一短単位と判断し，各々に分類語彙表番号を付与した。

人名については作業対象外としたが，地名や普通名詞を含む「名古屋タワープラザホール」「阪急グランドビル」のような固有名詞については，分割した短単位ごとの意味分類が可能と考え，「名古屋タワープラザホール」であれば，「名古屋（分類語彙表番号：1.2590）」「タワー（1.4410）」「プラザ（1.4700）」「ホール（1.2660）」のそれぞれに分類語彙表番号を付与する。

また，UniDic の語彙素に対応する語義として，たとえば「置く（語彙素番号：4811）」には「2.1513：固定・傾き・転倒」「2.1560：接近・接触・隔離」「2.3630：人事」の 3 つが結びつけられているが，例(1)では，担当作業者が分類語彙表を検索し，最適な意味分類として「2.1220：成立」を記入していた。同分類には「設立する」「設置する」「構える」などの見出し語が含まれる。このように，作業者が文脈上適切と判断した意味分類を新たに付与する場合もあった。

(1) マサチューセッツ州ケンブリッジに拠点を置くアカマイ・テクノロジーは，
（サンプル ID: PN1c_00004，下線は著者による）

このほか，たとえば分類語彙表では体の類にのみ番号のある「当時」「最近」「先月」などが文脈上副詞として用いられている[注7]など，「名詞-普通名詞-形状詞可能」「名詞-普通名詞-副詞可能」のような品詞の語については，そもそも分類語彙表上に分類が存在しない場合もある。そこで，分類語彙表に適切な意味分類がない場合には，相の類を新設するなど新たな分類番号を作成した。なお，新項目の設置や項目名の変更など，意味分類に関する操作は行わず，品詞の違いに際して類を追加するに留めた。

2.3 作業結果

自立語（のべ 182,166 語）について，文脈上適切な分類語彙表番号を人手により付与した。このうちのべ 19,438 語（自立語の 10.7%）については，新たに分類番号を記入した。すなわち，UniDic に掲載がなかった新語と，分類語彙表に掲載がなかったが文脈上現

表 2 「UniDic- 分類語彙表対応データ」の未登録語・意味分類

	1：体	2：用	3：相	4：他	総　計
語彙素番号なし 未登録語への付与	9040 （3651）	300 （187）	1279 （244）	81 （34）	10699 （4116）
語彙素番号あり 未登録語義の付与	2133 （647）	277 （158）	488 （194）	18 （7）	2917 （1006）
新規付与（総計）	11173 （4298）	577 （345）	1767 （438）	99 （41）	13616 （5122）

（のべ語数　（ ）内は異なり語数：自立語のみ）

れていた新たな意味分類が得られた（**表 2**）。

この結果は，UniDic 語彙素番号と分類語彙表番号を対応させたデータの拡充に利用することが可能である。また，分類語彙表になかった新規の意味分類番号が作成された（**表 3**）。この結果により，実データを反映させた分類語彙表の拡充が期待される。前節で例示した時間副詞関連の番号（3.16～）と UniDic で形状詞とされる「よう」「そう」などが新規番号として導入されている。

表 3　分類語彙表にない新規番号の導入

1：体	2：用	3：相
46（31）	7（4）	655（45）

（のべ語数　（ ）内は異なり語数）

3. 基礎情報

本稿で付与作業を行った 3 種類の媒体別に，分類語彙表番号における「類」「部門」「中項目」分布を示し，詳細な「分類項目」についても分布例を示す。

3.1 媒体別の類分布

まず，**表 4** に書籍（PB），雑誌（PM），新聞（PN）の媒体別の類分布を示す。主として「体（1）」は名詞，「用（2）」は動詞であるため，「体」「用」の類の分布は媒体別の品詞分布（『現代日本語書き言葉均衡コーパス』品詞構成表 ver.1.1）と同様の結果になっている。但し，修飾語の含まれる「相（3）」，接続詞や感動詞などが含まれる「その他（4）」は，新聞（PN）で少ない傾向が見られている。

表 4　媒体別の類分布

媒　体	1：体	2：用	3：相	4：他	分類対象外	総　計
PB（54,474 語）	55.01%	26.43%	13.18%	2.21%	3.17%	100.00%
PM（60,786 語）	62.81%	21.51%	11.43%	1.45%	2.80%	100.00%
PN（66,906 語）	73.53%	16.40%	6.65%	0.61%	2.81%	100.00%
総計（182,166 語）	64.42%	21.10%	10.20%	1.37%	2.91%	100.00%

（のべ語数　※「分類対象外」は外国語の機能語や固有名詞の一部など）

3.2 媒体別の部門分布

表 5 に媒体別の部門分布を示す。現代日本語書き言葉を構成する自立語を意味分類で見ると，「関係（.1）」が 46.5% を占め，次いで「活動（.3）」が 28.9%，「主体（.2）」が 13.3%，「自然（.5）」が 4.3%，「生産物（.4）」が 4.1% の順となっている。また，媒体別では，新聞（PN）では書籍（PB）や雑誌（PM）に比べ，生産物・自然物が少ない

傾向が確認できる。また，雑誌（PM）は新聞（PN）や書籍（PB）よりも主体が少なく関係が多いという傾向が見られる。

表 5　媒体別の部門分布

媒体	1：関係	2：主体	3：活動	4：生産物	5：自然	対象外	総　計
PB（54,474 語）	46.25%	12.07%	28.97%	4.32%	5.22%	3.17%	100.00%
PM（60,786 語）	47.68%	10.99%	28.41%	4.94%	5.18%	2.80%	100.00%
PN（66,906 語）	45.61%	16.45%	29.22%	3.08%	2.82%	2.81%	100.00%
総計（182,166 語）	46.49%	13.32%	28.88%	4.07%	4.33%	2.91%	100.00%

（のべ語数　※「分類対象外」は外国語の機能語や固有名詞の一部など）

3.3　媒体別の中項目分布

表 6 には同データの中項目分布を示す。

表 6　媒体別の中項目分布

中項目	PB	PM	PN	総　計
19：関係-量	9.72%	15.09%	18.96%	14.91%
30：活動-心	9.09%	8.01%	7.38%	8.10%
15：関係-作用	7.25%	7.29%	6.21%	6.88%
12：関係-存在	7.09%	5.30%	4.25%	5.45%
34：活動-行為	5.67%	5.34%	5.39%	5.45%
16：関係-時間	5.03%	4.33%	4.95%	4.77%
10：関係-事柄	6.67%	4.89%	2.77%	4.64%
31：活動-言語	4.72%	4.96%	3.72%	4.43%
11：関係-類	4.48%	4.28%	3.77%	4.15%
20：主体-人間	5.11%	3.31%	2.36%	3.50%
25：主体-公私	2.62%	1.84%	5.24%	3.32%

（総計上位（各媒体毎上位 6 種に網掛け））

前節で見た通り，部門は「関係（.1）」が約半数を占めるため，上位には「関係」部門の中項目が目立つが，「量（.19）」の分布が 14.9%と最も多く，「作用（.15）」が 6.9%，「存在（.12）」が 5.5%，「時間（.16）」が 4.8%，「事柄（.10）」が 4.6%という分布がある。媒体差を見ると，最上位の「量（.19）」において，新聞（PN）で特に多く雑誌（PM）が次ぐが，書籍（PB）では少ないという明らかな違いがある。「量（.19）」には数記号をはじめ助数詞，程度や全体部分などが含まれるためである。反対に，「存在（.12）」は「ある」「いる」をはじめ，出没や発生などの動作を含むため，書籍（PB）に多く新聞（PN）では少ない傾向となっている。各媒体の特徴として，書籍（PB）で「事柄（.10）」，雑誌（PM）で「言語（.31）」，新聞（PN）で「公私（.25）」が他媒体と異なる上位分類である

ことがわかる。「事柄（.10）」には事件のほか指示詞も含まれるため，書籍（PB）において特徴的な意味分類といえる。「言語（.31）」は引用の「という」や判断の副詞などを含み，雑誌（PM）に多く現れていた。「公私（.25）」は国や政治的区画，地名，法人などを含むため，新聞（PN）に特徴的であった。意味分類の分布を調査することにより，語彙的には表れにくい特徴が明らかとなる場合がある。たとえば，「言う」のみの頻度は書籍で最も高い（『現代日本語書き言葉均衡コーパス』短単位語彙表 ver.1.1）。しかし，「言う」の含まれる「言語（.31）」は，雑誌に特徴的な意味分類であった。こうした意味分類の活用により，新たな知見の得られる可能性がある。

3.4　媒体別の分類項目分布例

分類項目の分布を見ることで，意味ごとに媒体を調査することが可能となる。ここでは，例として中項目が「時間（.16）」の分類項目について**表 7** に示す。分類項目が多岐に渡るため，100 万語あたりの頻度とした。いずれの媒体でも，「時間的前後（.1670）」の頻度が高いものの，書籍（PB）は「時期・時刻（.1611）」が最も多く用いられており，雑誌（PM）は「場合（.1690）」が，新聞（PN）では「過去（.1642）」が次ぐという媒体別の特徴が見られる。「時期・時刻（.1611）」と「場合（.1690）」はいずれも「～とき」のような用例が含まれ，語彙素のみからは分析しにくいが，文脈から判断された意味情報が付与されたことで，意味的に区別が可能となった例である。また，このような分類項目の頻度調査によって，書籍（PB）と雑誌（PM）における「とき」の意味的な用法傾向の差なども確かめられるようになったといえる。なお，語彙素の内訳を見ると，新聞（PN）の「過去（.1642）」項目においては，「昨年（pmw1076 件）」が突出し，「過去」

表 7　媒体別の分類項目分布例

分類項目	PB	PM	PN	総　計
1670：時間的前後	6407	5462	5874	5896
1611：時機・時刻	7104	3866	3602	4737
1600：時間	5048	3866	3453	4068
1690：場合	4167	4392	2526	3640
1641：現在	3231	3619	3796	3568
1642：過去	2442	2731	3946	3091
1612：毎日・毎度	3084	2912	2436	2789
1660：新旧・遅速	2552	2517	2660	2580
1650：順序	2093	2089	3169	2487
1623：時代	2570	2468	2227	2410
1651：終始	1928	1497	1973	1801

（総計上位，調整頻度 pmw（媒体毎上位 7 種に網掛け））

の 27.3%を占めるが「去年（書籍のみ pmw18 件）」の例はない。また，「元（pmw942 件）」の頻度の高さが，書籍（PB，pmw92 件）や雑誌（PM，pmw115 件）に比して特徴的である。新聞が過去のことを扱う用例を多く含むことはもちろん，他媒体とは異なり決まった語句や接頭語が高頻度であることなどもわかる。分類項目の分布に特徴的な傾向を確かめることで，当該分類項目の語彙を調査し，詳細な傾向分析が可能になると考えられる。

4. まとめと展望

本データの作成と公開により，BCCWJ の一部が意味的な情報によって検索可能となった。従来用例の収集が困難であった意味分類を要する研究における利活用の可能性が期待される。今後，長単位と機能語についても意味情報を付与したデータを追加公開する予定である。また，現在進められている『日本語歴史コーパス』に対する分類語彙表アノテーションとの比較対照が可能となる。

データの公開は，https://github.com/masayu-a/BCCWJ-WLSP/ にてスタンドオフ形式のものを公開するほか，BCCWJ 有償版契約者には「中納言」のダウンロードサイトにてエクセル形式のものを公開する。

注 1　本稿は，以下の発表で途中経過を示したデータの完成と公開を報告するものである。
加藤祥・浅原正幸・山崎誠．2018.「『現代日本語書き言葉均衡コーパス』の新聞・書籍・雑誌データに対する分類語彙表番号付与」，『日本語学会 2018 年度秋季大会予稿集』，161-166.
S. Kato, M. Asahara, and M. Yamazaki. 2018. “Annotation of ‘Word List by Semantic Principles’ Labels for the Balanced Corpus of Contemporary Written Japanese” Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation.

注 2　分類語彙表番号の付与される短単位の機能語は助動詞と助詞の一部に限られるためである。なお，著者らは機能語について別途意味情報の付与を進めている。

注 3　優先順位は，https://pj.ninjal.ac.jp/corpus_center/anno/ に記載があり，本作業は優先順位 A・B グループの全サンプルを作業対象としている。

注 4　本データは，現代書き言葉 UniDic（https://unidic.ninjal.ac.jp/）としてダウンロードが可能である。

注 5　一般的な判断という観点から，クラウドソーシングを用いて募集した作業協力者に，判定の揺れた用例を提示した語義判定を依頼した。今後，追加調査によって集計結果を加味した修正版の公開があり得る。

注 6　長単位についても分類語彙表番号の付与作業を行った。「ていく」「てくる」をはじめ，「にとって」など，助動詞扱いとなるが短単位と異なる意味分類となる場合などは，機能語であっても分類語彙表番号を付与している。また，長単位より大きな単位（慣用句など）として分類語彙表番号がある場合には，メモとして番号を付与している。これらの情報の整理が完了すれば，あわせて公開予定である。

注 7　品詞性の判断については，BCCWJ コアに付与された人手による「名詞」「形状詞」など

の用法情報に従った。

【参考文献・資料】

小木曽智信・中村壮範．2014.「『現代日本語書き言葉均衡コーパス』形態論情報アノテーション支援システムの設計・実装・運用」,『自然言語処理』21(2), 301-332.

A. Kondo, M. Tanaka, and M. Asahara. 2018. “Alignment Table between UniDic and ‘Word List by Semantic Principles’”, Proceedings of Eighth Conference of Japanese Association for Digital Humanities.

F. Bond, T. Baldwin, R. Fothergill, and K. Uchimoto. 2012. “Japanese SemCor: A Sense-tagged Corpus of Japanese” in The 6th International Conference of the Global WordNet Association (GWC-2012)

K. Maekawa, M. Yamazaki, T. Ogiso, T. Maruyama, H. Ogura, W. Kashino, H. Koiso, M. Yamaguchi, M. Tanaka and Y. Den. 2014. “Balanced Corpus of Contemporary Written Japanese”, Language Resources and Evaluation, 48(2), 345-371.

M. Okumura, K. Shirai, K. Komiya and H. Yokono. 2011. “On SemEval-2010 Japanese WSD Task”,『自然言語処理』18(3), 293-307.

国立国語研究所（編）．2004.『分類語彙表増補改訂版データベース』https://pj.ninjal.ac.jp/corpus_center/archive.html#bunruidb

『現代日本語書き言葉均衡コーパス』短単位語彙表 ver.1.1，品詞構成表 ver.1.1

【謝辞】 本研究は，国立国語研究所コーパス開発センター共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」によるものです。本研究の一部は JSPS 科研費 17H00917，18H05521，19K00591，19K00655 の助成を受けました。

――かとう さち　国立国語研究所プロジェクト非常勤研究員――
――あさはら まさゆき　国立国語研究所教授――
――やまざき まこと　国立国語研究所教授――

（2019 年 1 月 31 日　第 1 稿受理）
（2019 年 4 月 24 日　最終稿受理）